Social Media Neighborhood Guides

A PhD Thesis Proposal by Dan Tasse

Committee:
Jason I. Hong (Chair)
Jodi Forlizzi
Niki Kittur
Judd Antin

Abstract

Modern tourists visiting new cities are not content to simply stay in a hotel downtown and see famous sights. They want to get out into the neighborhoods of the city that they are visiting and understand more of the city's culture and everyday life. However, current guides remained focused on statistics and points, so tourists are unable to understand and find neighborhoods they would enjoy.

I propose to build neighborhood guides based on social media posts to help people understand neighborhoods. These guides will have two parts: first, they will allow comparison between neighborhoods in a new city and neighborhoods they know; second, they will add context so travelers can understand why the neighborhoods are similar. These will enable people to understand how different neighborhoods feel, and contribute to our understanding of the city as a whole. Their effectiveness will be evaluated through quantitative studies of the comparisons and qualitative studies of the site as a whole.

This thesis will provide three research contributions. First, it will provide evidence that social media can help us understand cities better than simple demographics. Second, it will show how well social media reflects neighborhoods, and what aspects are best represented. Finally, it will contribute to our knowledge of tourist information search by the development of a five dimensional model.

Table of Contents

Abstract 2

Chapter 1: Introduction 4

Chapter 2: Background/Related Work 7
- **Changes in Urban Tourism** 7
- **Computational Tourist Experience Recommendations** 7
- **Summarizing Geographical Social Media Posts** 8

Chapter 3: Completed Work 10
- **Analyzing Tweets To Find Where Tweeters Live** 10
  - Data Collection 10
  - Methods for Finding Home 11
  - Results 12
- **Using Tweets to Characterize Locations** 14
  - Creating the Twitter Neighborhood TF-IDF Map 14
  - Results 15
- **Understanding Travelers' Needs** 16
  - Method 16
  - Finding 1: People use heuristics when searching, if possible 17
  - Finding 2: If no heuristics are available, people attempt to satisfy five dimensions. 18
  - Finding 3: Current search tools do not adequately investigate those five dimensions. 21

Chapter 4: Proposed Work 23
- **Neighborhood Comparison Algorithm** 24
  - Safety and Room for Everyone: US Census Demographics and Crime Statistics 24
  - Aesthetics: Flickr autotags 24
  - Serendipity: Walkscore and Transitscore 25
  - Ideal Everyday: Third Places from Yelp 25
  - Authenticity: Tweets 26
- **Context for Neighborhood Comparisons** 27
- **Evaluation** 28
  - Are the neighborhood comparisons "right"? 28
  - Is this guide useful? Does it reflect the city accurately? 29
- **Contributions** 29

Chapter 5: Schedule 31

Acknowledgements 31

References 31

Chapter 1: Introduction

Many of the readers of this thesis proposal will soon be heading to Cologne, Germany for the ICWSM 2016 conference. Let's say you are one of them. You may have some extra time before or after the conference to relax a bit and enjoy Cologne. You will probably stop in the famous Cologne Cathedral, see the Hohenzollern Bridge, and maybe even visit the Museum of Chocolate if you have a sweet tooth. But what about the more everyday Cologne, the "non-touristy" side, the part that is great for the people who live there? Surely Germany's fourth largest city has more to offer than a grand cathedral and a few other tourist spots. You may be interested in staying in an AirBnB too, to save some money and meet some locals… but where? What neighborhood is close enough to the conference but also intriguing and friendly enough to stay in?

This is a specific case of a general problem: we need new ways to understand cities and neighborhoods. As more people move to cities throughout the 21st century, quickly understanding how places feel will become more and more important. People moving will need to know what neighborhood they would feel at home in, business owners will need to know where to expand and market, and city planners will need to know how to allocate services and zone districts.

Travelers have a unique set of information needs, though, because they are new to a place and do not have time to build up local knowledge from experience. More than ever before, too, they *want* this local knowledge; they want to experience "everyday life" in a city and "do what the locals do." Unlike the sun-and-sand tourists of two generations ago or the cultural-site-visiting tourists of last generation, today's tourists want to curate and create their own experience. And more than ever, platforms like AirBnB and Couchsurfing help them do so by staying in local neighborhoods instead of central tourist districts.

Tools that are available to address these information needs all fall short. Traditional guidebooks from Fodor's, Frommer's, and Lonely Planet give people information about those central tourist districts and sights to see. Yelp and Foursquare give people information about the businesses, the bars and restaurants and locksmiths, in an area, but travelers can't understand how the whole neighborhood feels just from that. Cities gather statistics – and indeed, are releasing open data more than ever before – but numbers also fail to convey a neighborhood's culture. Finally, occasionally travelers can learn local vernacular descriptions, but these are often shallow. For example, "Lawrenceville is the cool neighborhood" or "South Side is the party neighborhood."

However, thanks to public geotagged social media posts, we have enough localized information to help inform these travelers. Travelers want to stay in places that satisfy certain cultural and aesthetic criteria that are better reflected in Tweets and photos than in statistics and lists of tourist sites.

I propose to build a web-based neighborhood guide for travelers, based on social media posts, that will help travelers find neighborhoods they will enjoy staying in and spending time in. I will do this by comparing neighborhoods to neighborhoods in a city they already know, to use people's existing understanding of neighborhoods to scaffold their understanding a of a new city. The algorithm for comparing neighborhoods will be based on five dimensions derived from existing research and from formative interviews. The neighborhood guide will also, importantly, contain ways to understand why two neighborhoods are similar: extra context in the form of photos, text excerpts, or relevant statistics.

Using this guide, tourists will be able to find places to stay and places to spend time more easily. This will make it easier and more fun to travel to big cities, but also more fun to travel to mid-sized or small cities that currently do not get as much tourist attention. With international travel destinations like Paris and Venice losing character due to a deluge of tourists and smaller but worthwhile cities like Cleveland and Atlanta needing ways to attract investment, this could benefit the entire tourism industry. Furthermore, this guide could be useful for planners beyond travelers, especially for neighborhoods that are growing or developing and want to be like other more popular neighborhoods.

Point-based guides and statistics can only go so far to help us understand crucial aspects of a city's culture, and travelers nowadays want to know that culture more and more. This tool will help people deepen their understanding of cities, and help researchers learn about cities as well.

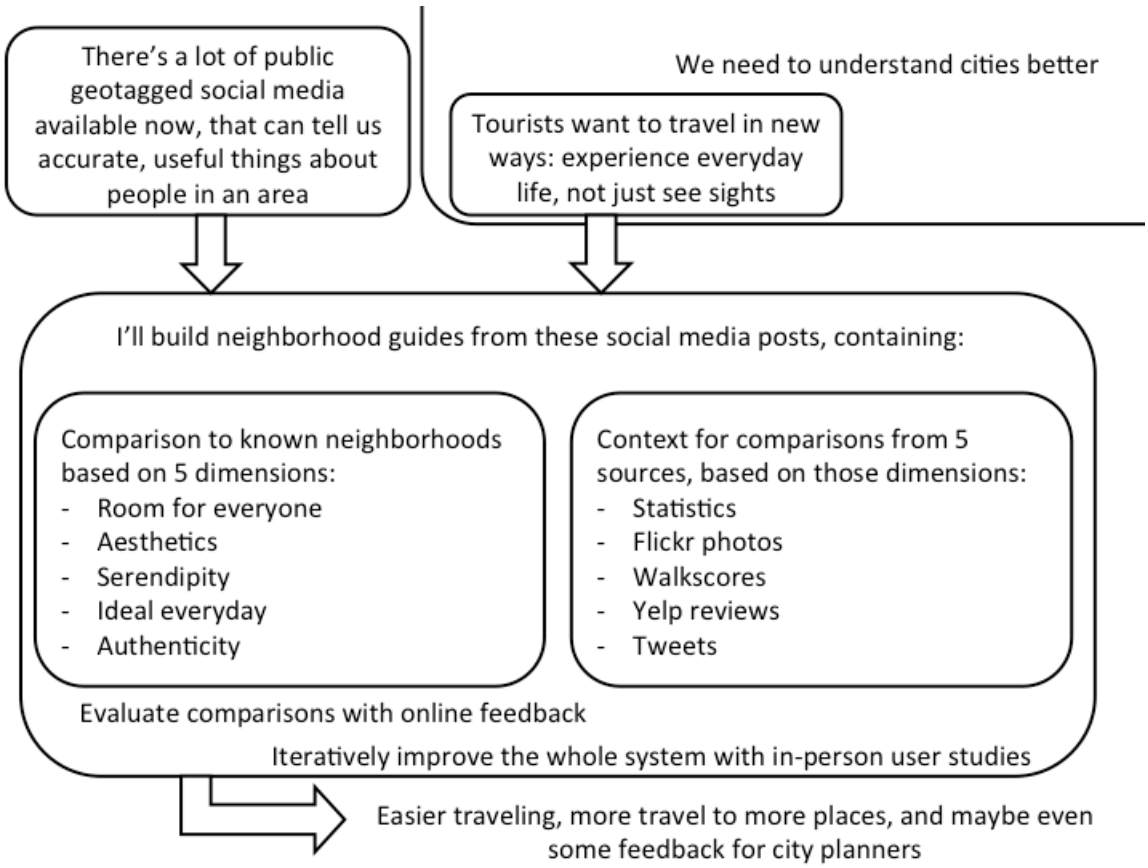

Figure 1: A conceptual overview of the work in this thesis proposal

Chapter 2: Background/Related Work

Related work falls into two main categories. Recent work in urban tourism provides motivation for new tools to navigate neighborhoods. At the same time, work in computational recommendation of tourist experiences suggests one possible avenue to build tools for travelers. However, I will explain why that trajectory is unsatisfactory, and describe work in summarizing geotagged social media that offers more promise in helping us build useful tools.

Changes in Urban Tourism

While urban tourism was not a focus of early tourism research, it has recently become a growing field [11]. Travel in previous decades had meant traveling to beaches, beautiful natural sites, or resort towns, but in recent years urban tourism is the fastest growing segment of the tourism market [4]. The character of urban tourism is changing as well as the volume: new urban tourists want to "experience and feel a part of everyday life." [27] Furthermore, they seek to have an active hand in co-creating the experiences, rather than passively paying for and absorbing an experience [2]. Lists of sights to see and experiences to buy no longer suffice.

In addition, when modern tourists travel to a city, they are often looking for an authentic experience of that city, rather than a manufactured diversion. The search for authenticity in tourism has a long history dating back at least to the 1970s [19], but recent developments have aided this search in new ways, particularly with regard to lodging. Because hotels historically clustered in a few areas of cities, like downtown and near airports, they cannot show travelers all the sides of a city they may want to see, so travelers are turning to alternatives. The peer-to-peer lodging rental site AirBnB, for example, has become a popular, and more "authentic", way for travelers to rent rooms in residential parts of town [43, 49]. Similarly, Couchsurfing allows users to stay with locals for free (often on their spare couch, hence the name) [49]. As urban tourists change from "mass tourists" to "cultural" and "creative" tourists [41], "mass" lodging no longer suffices either.

New urban tourists want to stay in interesting residential neighborhoods and spend time "wandering about", "taking in the city", and "getting among the people" [2]. To do this, they need guides to areas, not specific venues. Urban tourism, unlike other forms of tourism like "sun and sand" tourism, depends on the serendipity and spontaneity that results from getting to know neighborhoods, and on the individual's ability to co-create their experience. Current tools help people discover points, not overall pictures of parts of the cities.

Computational Tourist Experience Recommendations

While travelers have been changing, plenty of work has gone into addressing exactly the problem of recommending things for tourists to do. Work in this vein includes recommendations of restaurants [14], shops [44], travel routes [20, 32], attractions and points of interest [12], and destinations [13]. These all use social media and user-generated content such as user locations, so continuing in this vein seems like a logical choice. In addition, sites like Yelp and Foursquare have dozens of user reviews, so aggregating reviews and recommending the most highly-rated spots seems like a natural solution.

However, this approach has three shortcomings. First, people need to know why they are recommended each place. It would be rare for tourists to set out on a trip solely because an algorithm recommended it. Second, they solve problems that are already solved by Yelp and Foursquare: finding a restaurant or a point of interest by consulting one of these guides is easy. Finally, these works neglect the changes in urban tourism discussed recently. A recommendation algorithm will likely push more people to the top destinations, which then become overcrowded and no longer as enjoyable. Instead, we need guides to let people explore places on their own time and create their own connections to them.

Summarizing Geographical Social Media Posts

While the recommendation of tourist places work has been going on, a separate set of researchers has been investigating public geotagged social media posts: public photographs and text posts with latitude and longitude tags attached. Photo-sharing sites, particularly Flickr, have been well studied, due to the volume and richness of their posts. Some of this research has been driven by practical concerns, like the need to show photos on a map. Toyama et al [46] developed techniques including thumbnails, point markers, and isopleths to show how many photos existed on a map at a place before settling on a binning approach they call “media dots.” However, these displays only show the number of photos, not their contents, so a series of other projects worked on summarizing photo content as well as density.

Some of this research works on finding a subset of photos that is representative of a larger set. Jaffe et al [15] addressed the problem of summarizing photo content by finding a subset of photos that would accurately summarize a larger photo set. They did this by clustering all of the photos and then ranking the clusters based on five criteria: tag distinguishability, photographer-distinguishability, density, image qualities, and arbitrary relevance factor (such as a search query). Kennedy et al [17] further developed the ability to find the “most representative” image from a set of photos using computer vision features such as SIFT. Crandall et al [6] did the same: finding the top N “interesting” places in each city and a “canonical” photo from each.

Besides investigating photo contents, researchers have investigated ways to summarize the textual tags that users add to their photos. Ahern et al [1] and Jaffe et al [15] describe the World Explorer/Tag Maps project, which summarized a series

of photo tags into "representative tags" for a region. Kennedy and Rattenbury expanded this to describe semantics of places and events [17, 38], while Kafsi et al further expanded it to understand which tags are locally relevant, which are city-level, and which are country-level [16].

Summarizing textual content, like tweets, is somewhat easier because there is less total information, so one can use a simple method like a word cloud (at least as a supplementary tool) to get a sense of a large corpus of words [29]. More intelligent methods have been used for tweets, for tasks like event detection [19]. Importantly for neighborhoods, though, Hao et al approach high-level neighborhood modeling in another interesting manner, creating Location-Topic Models based on what users write in travelogues [13].

These algorithms, therefore, are now part of our toolbox: we have ways to summarize photo content, photo tags, and plain-text microblog posts.

However, higher-level abstractions can be useful too. Sometimes someone has a lot of data of one type and wants a simple summary of that data, but often more abstract representations are more useful because we can understand them better. The Location-Topic Model is one of these higher-level tools; two more that are worth discussing include neighborhood boundary finding and neighborhood comparison.

The flow of people throughout neighborhoods is often not reflected in the official neighborhood divisions, but recent work has been able to find boundaries based on human behavior such as Foursquare checkins [7, 51] or tweets [47]. This can reveal aspects of neighborhood life that is otherwise hidden, such as a neighborhood that contains two mostly-separate social sub-neighborhoods.

Finally, neighborhood comparison [22] offers a way for people to understand neighborhoods in a new city based on neighborhoods that they already know. This can help people talk about imprecise or unnamed characteristics of neighborhoods – they may not know what they like about their home neighborhood, but they know that they want to find someplace like it. This will be a key part of the neighborhood guide I propose to develop in Chapter 4.

The work in this chapter presents three types of work:

- Motivation for a new kind of travel guide, because traveler desires are changing and current guides are not serving their needs.
- One approach that, while useful and innovative, will not satisfy this new generation of travelers.
- Tools that summarize social media data, which provide both evidence that social media can be useful to describe places, and some tools that we can reuse for future work.

Chapter 3: Completed Work

To build better neighborhood guides based on social media, I began by investigating social media and what it can tell us about the people posting it. I looked at where Twitter users live and found that we can tell where about 80% of Twitter users live, which means that we can accurately use their tweets to tell us about their neighborhoods. I then built a preliminary neighborhood guide based on tweets, which revealed a few interesting findings about Pittsburgh's neighborhoods. Finally, I ran qualitative interviews with 24 participants to understand what people are actually looking for in these guides. I will describe these three projects in this chapter.

Analyzing Tweets To Find Where Tweeters Live

Our first study involved an investigation into geotagged tweets and how well we could find the homes of the tweeters[1]. This is a crucial first step in making use of social media data; without this context, it is not clear whether a geotagged tweet comes from someone who is very familiar with the place or someone who just visited once. Previous work has focused on localizing individual tweets [26, 37] and finding the homes of social media users [25, 36], but no work has specifically focused on finding tweeters' homes based on their geotagged tweets.
We did this by gathering tweets, asking users for their home locations, and then testing various algorithms to see how accurately each one found users' home locations.

Data Collection

To build a ground truth data set, we began by collecting 3.3 million geotagged tweets via Twitter's public streaming API. This API allows a developer to listen for new tweets that match a geographic parameter in near real time, so we chose to stream all tweets geotagged within 0.2 degrees latitude and longitude from the center of Pittsburgh. The rectangle we selected had corners at (40.241667, -80.2) and (40.641667, -79.8), and we collected tweets from January 2014 to January 2015. Following other work [30], we can assume that if our sample is less than about 1% of all tweets, we collected the vast majority of geotagged tweets in the region.
Near the end of that year, we used our data set of streamed geotagged tweets to compile a list of the 4119 most prolific tweeters for analysis, in order to ensure that our participants had enough geotagged tweets to analyze. We recruited these prolific tweeters to take a survey by tweeting a link to them. Our survey asked seven questions: their age, gender, home address, length of time they had lived there, work address, standard commute mode, and any other places they spend a lot of time. Respondents were paid with a $5 Amazon.com gift card via email. We received 195 responses.

[1] The work described in this section will be appearing in the proceedings of ICWSM

For each of our 195 users, we used the non-streaming Twitter API to gather that user's previous 3,200 tweets (the maximum number allowed by Twitter). We added any geotagged tweets that occurred outside of Pittsburgh to our data set. The data collection and survey process were approved by our university's Institutional Review Board.

Our final data set consisted of 146,852 geotagged tweets from 195 users, who had a median of 533 geotagged tweets (mean=753, min=15, 1st quartile=271, 3rd quartile=1050, max=3639). These represented a subset of all of their tweets; the median percent geotagged was 41.1% (mean=46.2%, min=2.3%, 1st quartile=25.1%, 3rd quartile=61.6%, max=100%).

One notable surprise in our data set was that we had many young participants (mean=26.9, median=22). This may be because Twitter is most popular with younger users [8] or because younger users felt more comfortable revealing their personal information on our survey. Many of these young 18-22 year old participants were students who had multiple "homes": they lived at their family home (often outside of Pittsburgh) during the summer and at their campus home (in Pittsburgh) during the school year. Because the school year lasts 8 months or more, we asked them on the survey for their campus home, but many of them still put their family home. As a result, we manually edited 19 students' "home" addresses to be their campus addresses, based on inspection of their tweets showing places where they talked about being "home" near a university.

Methods for Finding Home

In this section, we present a systematic evaluation of several algorithms for finding users' homes. In this paper, by "finding users' homes", we mean predicting a latitude-longitude point that is as close as possible to the geocoded address that they provided. We do not do reverse geocoding to find a street address.

Baseline (Mode of Geotagged Tweets)

As a trivial baseline, we binned tweets by rounding each tweet to the nearest 0.01 degree of latitude and longitude, then predicted that the bin with the most tweets (i.e. the mode) was the user's home location.

Last Destination, Weighted Median, Largest Cluster

Krumm [18] found people's homes based on GPS traces of their cars. We re-implemented three of his methods:

- Last Destination, where we take the median of the latitude and longitude of all points that are the last coordinate pair of the day (where a day ends at 3:00 AM)
- Weighted Median, where each point is weighted by the time until the next point
- Largest Cluster, using the scikit-learn [35] implementation of agglomerative clustering on all tweet locations

Grid Search

We binned tweets as in the Mode algorithm, but did so recursively, as in [5]. First we rounded tweets to the nearest whole number degree and discarded all tweets outside the most common bin. We repeated this rounding to the nearest 0.1 degree, the nearest 0.01 degree, the nearest 0.001 degree, and the nearest 0.0001, predicting the latter as their home.

Multi-level DBSCAN

To cluster points in a more principled way, we used the DBSCAN algorithm [10], as implemented in the scikit-learn library [35], to cluster tweets into clusters of different sizes. We set the Eps parameter (maximum distance between two samples in the same neighborhood) to be 0.2 degrees (latitude/longitude) for "city-level" clusters, 0.005 degrees for "neighborhood-level" clusters, and 0.0005 degrees for "building-level" clusters[2].

For each user, we chose the city-level cluster with the most tweets, then chose the neighborhood-level cluster with the most tweets, then the building-level cluster with the most tweets. We guessed that the centroid of the building-level cluster was the user's home location.

Grid Search Without Cross-posts

Given the similar accuracy of grid search and DBSCAN, we returned to grid search with a revised data set. We realized that 10.4% of our Twitter data set (15,261 of 146,852 tweets) were cross-posts from social apps. These apps include (in descending order of frequency) Foursquare/Swarm, Instagram, Untappd, Path, Camera on iOS, Spotify, MLB.com At the Ballpark, Frontback, Wordpress.com, Klout, LivingSocial, Sportacular, and MySpace. In each of these social apps, tweeting was a byproduct of another action (as op- posed to Twitter clients such as Tweetdeck and Tweetcaster). Furthermore, most of these are intended to be used outside the home. Therefore, they cannot help (and indeed would hurt) any home-finding algorithm. We removed them from the data set and performed grid search and DBSCAN again.

We then reasoned that nighttime tweets (from 8:00PM to 6:00AM) would be more predictive of home location than daytime tweets, so we removed daytime tweets and ran our algorithms again. This removed 77,122 of our tweets, leaving us with 54,469 tweets. We found the highest accuracy removing both of these data sets.

Results

Results are shown in Table 1.

Algorithm	Cross-posts	Night	Median	% of users	% of users	% of users

[2] Of course, "distance" does not make sense in terms of degrees longitude, because the length of a degree of longitude varies based on the latitude. However, because most of the points we considered were at similar latitude, we accepted this inaccuracy in order to test the method.

	removed	only	error	within 100m	within 1km	within 5km
Mode			553m	1.5	63.1	79.0
Grid Search			57m	54.4	73.3	86.7
Grid Search	✓		54m	56.2	76.8	88.1
Grid Search		✓	51m	56.2	77.3	87.6
Grid Search	✓	✓	49m	56.9	79.0	88.2
Multi-level DBSCAN			75m	52.8	72.3	87.2
Multi-level DBSCAN	✓	✓	75m	52.3	74.4	87.2
Last Destination			350m	40.5	66.7	85.6
Last Destination	✓	✓	520m	33.3	64.1	82.6
Weighted Median	✓	✓	400m	40.5	65.6	79.0
Largest Cluster	✓	✓	362m	33.8	69.7	87.1

Table 1: Results for each algorithm trying to predict each user's home. Best results are in bold. Results for Weighted Median and Largest Cluster without cross-posts and daytime posts removed were significantly worse, so we do not present them here.

These results show that, if you take away cross-posts and daytime posts, simple grid search shows where people live. Furthermore, these users do not need to have many tweets in order to be easily localizable, as shown by Table 2.

Last N Tweets	Median error	% of users within 100m	% of users within 1km	% of users within 5km
1	245m	44.6	61.7	74.1
5	84m	51.3	66.3	76.2
10	62m	58.0	75.1	81.9
100	65m	56.0	74.6	86.0
1000	51m	57.0	79.3	88.6

Table 2: Results using grid search on the most recent N non-crosspost non-daytime tweets for each user, for various values of N. Using more tweets allows better prediction, but prediction is remarkably good with as few as 10 tweets.

In summary, in this work we showed that it is relatively easy to find people's homes based on their geotagged tweets. We even improved the accuracy substantially over baseline.

Doing this analysis also helped me clarify the promise of geotagged social media. Finding these people's homes makes for useful analysis, but only for the 1% of tweeters who geotag their posts. Therefore, it is unlikely to be possible to locate any given person. However, by finding these people's homes, we were able to find a lot of people who live in a certain area, so we could use their tweets to figure out what (some percentage of) locals are saying in an area. This could really help us

characterize different places, which led us to the Twitter Neighborhoods TF-IDF Map, which we will explain in the next section.

Using Tweets to Characterize Locations

Based on these insights from our attempts at home finding, Jennifer Chou (an undergrad that I mentored) and I created our first attempt at a neighborhood guide, the Twitter Neighborhood TF-IDF Map (Figure 2).

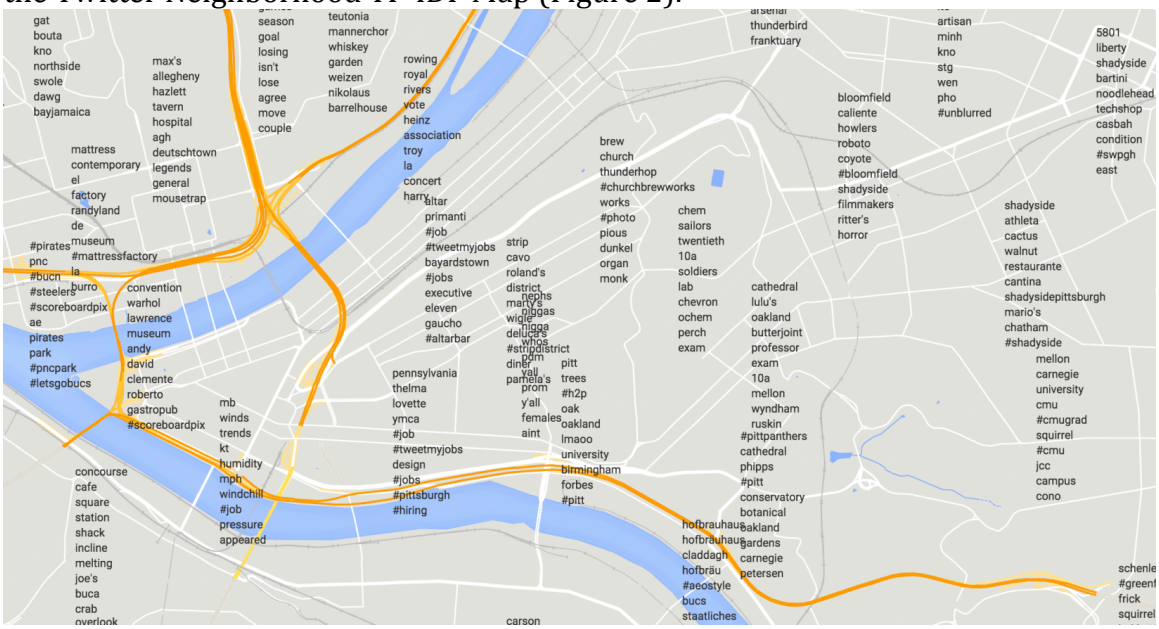

Figure 2: Most frequently tweeted words in each Pittsburgh neighborhood

This map shows which terms are used more often in one neighborhood than in others. For example, the Pittsburgh Pirates baseball team's hashtag `#pirates` is tweeted in many neighborhoods in Pittsburgh, but it is most often used near the baseball stadium.

Analyzing tweets gives us a unique window into these neighborhoods. While governments collect demographic data, that only tells quantitative facts: deep but narrow. Analyzing tweets can give a qualitative picture of a neighborhood: not quite what people in that neighborhood think or care about, but at least what those people say. Twitter users are admittedly a small sample of people in an area, but our work here suggests that those people can tell us useful insights about their neighborhoods.

Creating the Twitter Neighborhood TF-IDF Map

To create this map, we used the same set of tweets that we had gathered for the previous project: 3.3 million geotagged tweets in the Pittsburgh area. We then assigned each tweet a neighborhood based on its location, using neighborhood

boundaries from the Western Pennsylvania Regional Data Center[3]. We then applied a variant of TF-IDF to each word to find which words are the most indicative of each neighborhood. For our purposes here, TF, or term frequency, represents the number of times that word appears in tweets in that neighborhood; DF, or document frequency, represents the number of times that word appears in other neighborhoods. To find the TF-IDF score for each word in each neighborhood, we divide its term frequency by its document frequency. We then removed all words that were tweeted by fewer than 5 people, to reduce spam. (Other corrections, such as the TF-IDF-UF score used in [1], also seem promising.)

Results

Findings for this project were anecdotal, but suggested that this was a promising direction to continue. In looking at the map, we found obvious results, like where each stadium was, but we also found less obvious local references. For example, we learned that the 10a shuttle on the University of Pittsburgh campus was a frequent topic of conversation and jokes (Figure 3).

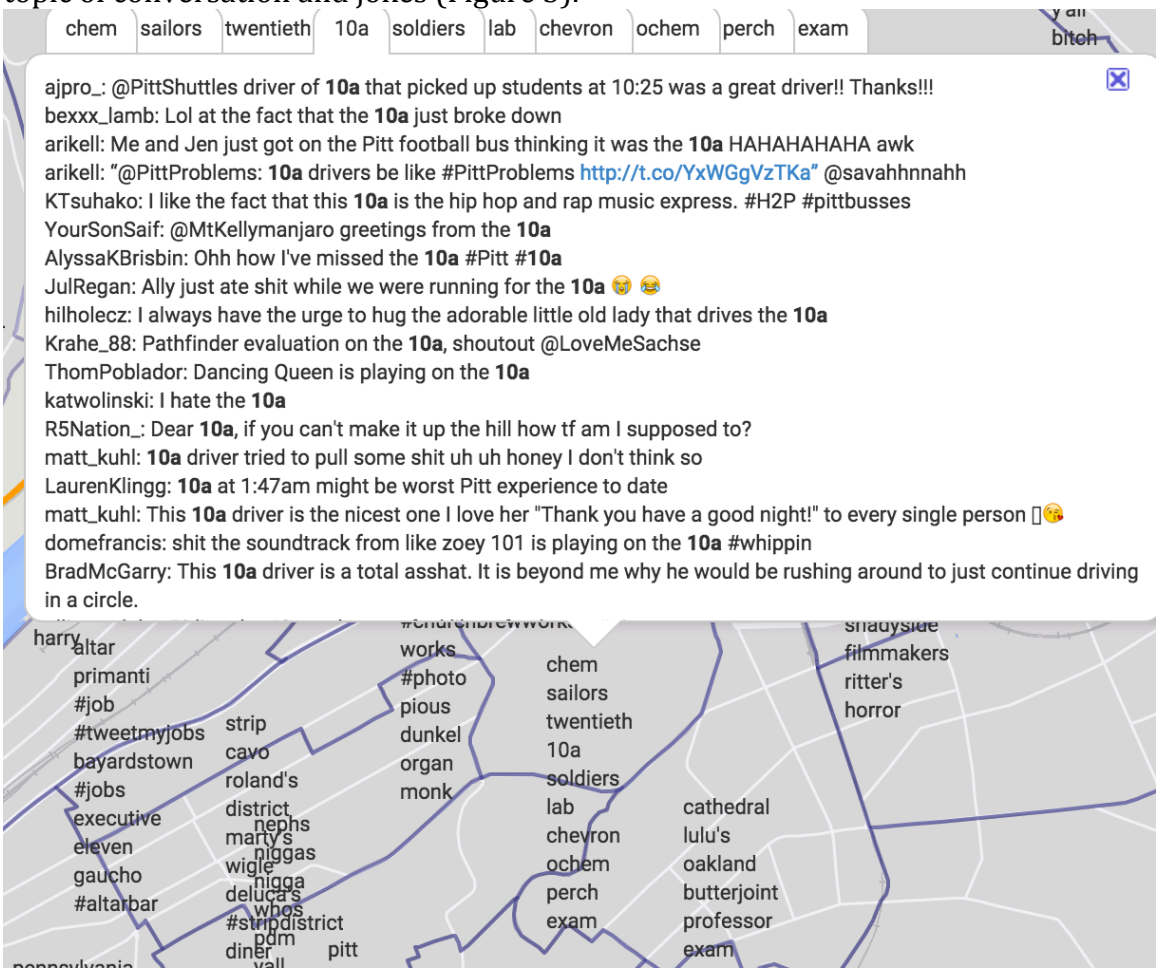

Figure 3: Tweets referencing the 10a bus

[3] https://data.wprdc.org/dataset/pittsburgh-neighborhoods770b7

In informal talks with friends and other students, we found lots of enthusiasm for tools like this, especially for cities that they do not know as well as Pittsburgh. However, we also found the results a bit lacking. Some of the findings were obvious, some were simply names of locations, and many failed to give users a good sense of what the places felt like. People that we talked to wanted richer insights into the neighborhoods, not just a few key words or locations.

To understand what they meant by "feel" and "richer insights", I embarked on a qualitative study to understand how people got to know neighborhoods and cities when they moved and traveled.

Understanding Travelers' Needs

To understand how people make sense of cities and neighborhoods now, I conducted interviews with recent movers and travelers. I knew I could use social media to give people some sense of a neighborhood, but I wanted to make sure I was creating something useful that filled a particular need. As a result, I focused on the following research questions:

- What do people want to know about neighborhoods when they're moving?
- What do people want to know about neighborhoods when they're traveling?
- What do people wish travelers and movers knew about their neighborhood?
- What parts of public social media will be most useful?

Method

I recruited 17 participants in Pittsburgh who all recently traveled or moved by posting our study on Reddit, Craigslist, and Facebook. We asked them to describe their search process and their experience finding a neighborhood to stay or live. We then showed them printed pages about the neighborhoods they moved and traveled to and from: popular Twitter words from the TF-IDF map, Flickr photos obtained by searching the neighborhood names on Twitter, the top 10 most popular venues on Yelp, and market research and statistics from ESRI's Tapestry guide[4]. We asked them to create two one-page guides (one for the neighborhood they moved/traveled to, and one for the neighborhood they moved/traveled from) by cutting and taping these materials, and writing or drawing in anything that was missing. This was meant as an elicitation exercise to get them thinking about these neighborhoods in more depth. CMU's Institutional Review Board approved this study.

After these 17 interviews, of which 7 involved recent travelers and 10 involved recent movers, I realized one recurring issue: social media is much better poised to help travelers than movers. Movers care about many factors in addition to the neighborhood: the house or apartment itself, the cost of rent or a mortgage, the proximity to an existing job, and the local schools. Some of their concerns still echo

[4] http://www.esri.com/data/esri_data/ziptapestry

the travelers' concerns, so I retained their data, but I reoriented the project to focus on travelers.

I recruited seven more recent travelers in San Francisco, bringing the total to 24 participants. For this second group, I did the same interview, but focused more on factors that seemed relevant in the first one: safety, liveliness, diversity, and aesthetics. I also only recruited travelers for the second group. I did not bring the printouts or ask participants to create flyers like I did for the first group, because the complication did not provide much more value or insight.

I will refer to the original 17 interviewees as A1-A17, and the next seven as B1-B7. These participants were young: all in their 20s and 30s except for two. Eight were students, while the rest were mostly professionals. Interviews were conducted in cafés or other public places near them for convenience and to get them thinking about their neighborhoods. B5 and B6, a dating couple, interviewed together; all the rest were done separately.

Because the interviews occurred in public places, I could not record the interviews, but I took plentiful notes to capture important points as well as possible. After finishing each batch of interviews, I analyzed the data iteratively, using an open coding approach inspired by grounded theory to allow insights to emerge from the data.

These interviews revealed a lot about this group's travel and moving motivations, what they hope to learn about neighborhoods, and where they decide to stay, as well as a few interesting tensions that arise when they make those decisions.

Finding 1: People use heuristics when searching, if possible

First, a number of conditions may cause travelers to do very little research before choosing where to stay. If someone already has a place to stay, they will likely take that. B2 described this as a "bird in the hand" situation, and said it occurred a lot when Couchsurfing: finding a local who's willing to host him for free can be difficult, so he will usually accept, regardless of circumstances.

If a traveler has social or other constraints, such as friends or family to visit or an event to attend, they usually consider tourism secondary and stay somewhere nice near that constraint. B5 and B6 described going to the X Games, an extreme sports event, in Aspen, Colorado: they spent most of their time watching events, so they simply wanted to stay near the games. Similarly, B4 described visiting Scottsdale on personal business (he declined to describe it further), which led to him staying in the Fashion Square district. He found it rather unpleasant, and had trouble getting around, but he needed to be near there.

Finally, budget constraints would often short-circuit the lodging search. B5 and B6 described another trip, when they went to Seattle but wanted to pick the cheapest lodging possible. This ended up being the Green Tortoise Hostel downtown, and

since they had stayed in another Green Tortoise elsewhere, they decided it would work. B3 also described a road trip where he simply looked up a place to stay while on the road each day, only wanting something simple, clean, and cheap.

Finding 2: If no heuristics are available, people attempt to satisfy five dimensions.

Most of the participants described trips where they did not use any of these heuristics, and instead wanted to satisfy five different dimensions: Safety and Room for Everyone, Aesthetic Appeal, Opportunity for Serendipity, The Ideal Everyday, and Authenticity. I will describe them in turn.

Dimension 1: Safety and Room For Everyone

Everyone wanted to be safe. The meaning of safety varied slightly depending on location; usually it included crime, but A1, A15, A17, and B4 all mentioned fear of bedbugs when traveling to New York. However, when asked if safety was always an upside, most participants declined. A6 described spending one night in Churchill Gardens, a posh part of London, but then moving on to somewhat simpler Clerkenwell. Often the safest spaces are also the most expensive, and because they are so expensive, only a homogenous set of wealthy people can live there.

Everyone who spoke of diversity considered it a virtue. They described enjoying markets (B1 and B6) and train stations (B1), as they are places where lots of different people meet. When asked why, they often mentioned gentrification. I took care not to introduce the term myself, but seven participants brought it up independently. A loss of diversity makes a place less fun (B4) but also brings about changes that make their existence in a place uncomfortable (A3, A10, A15, B1), because they're not sure if their existence there displaces other people. This principle was best articulated by A1 and B1, who both used the phrase "room for everyone": they want to see a place that contains people of diverse ages, races, and income levels.

Dimension 2: Aesthetic Appeal

Aesthetic appeal in many forms is one of the main incentives for people to travel, and one of the main influences on the overall feeling of a trip. By "aesthetic appeal", I am referring to anything about the senses: participants mentioned visual, auditory, and gustatory appeal particularly, and occasionally smell. Some preferences were universal, such as enjoyment of nature and avoidance of loud places while sleeping. Others were personal: A10 described her neighborhood as a burgeoning urban agricultural area, while B4 described the city of Pittsburgh as a "concrete jungle." Many participants described suburbs as "boring", but A7 described one suburb as his "perspective of what country living should be."

Dimension 3: Opportunity for Serendipity

Lodging seekers used a heuristic if they had one place to travel to (they would simply stay near that place), but travelers without a direct goal still valued convenience. What does "convenience" mean when one doesn't have a goal? It depends on the person and the city, but participants talked about "being mobile"

(B2), "being in the middle of stuff" (B4), "having stuff around" (A15), or being "where everything is" (A9). Even so, this is not very descriptive.

The final two interviews, though, helped elucidate this point. B6 talked about visiting New Orleans and stumbling across parades put together by local Native American groups, which she unexpectedly enjoyed. B7 described staying in the neighborhood of Itzimna, Merida, Mexico, which was a short walk from the tourist center of downtown. She enjoyed the walk downtown because it enabled her to discover more than if she were actually staying downtown. Both of these cases support the claim that "convenience" is more than just quick travel time to sights; it's about opportunities to discover these unknown gems. These opportunities appear more in a dense urban environment full of local businesses and walkable neighborhoods.

Walkability deserves extra attention here as a key enabler of serendipity. Traveling by car involves difficulties such as driving in a new city, covering unfamiliar terrain, and parking, as B5 mentioned. However, being stuck in a car-centric environment without a car is unpleasant, as in B4's trip to Scottsdale. Therefore, logistically, traveling is easier when one can just walk or use public transportation. In addition, exploring is easier when a stop into a store or café involves just walking in, not noticing it, finding parking, and walking in. When one can explore more, one can encounter more delightful, fortuitous experiences.

Dimension 4: The Ideal Everyday

One recurring theme was described as "taking in the city life" (B1), seeing "what people actually do here" (A9), "kind of get[ting] a feel" of the city (A6), and even "play[ing] the game of, what if we lived here" (A17). This echoes a trend towards travelers using the everyday as a way to create their experience, for the travel experience to be less about what they are consuming and more about what they are becoming [27]. Most participants (except A16) were not traveling in order to find a place to move to, but they still enjoyed pretending to do so.

When pressed, though, interviewees did not actually want their travel experiences to be about the real "everyday." Everyday life involves work, chores, and errands that most people do not enjoy, wherever they are. For example, asked if she would be interested to see everyday life in the Financial District of San Francisco, B1 replied no, the Financial District isn't the kind of "everyday" she's looking for (though clearly it is an integral part of many people's everyday lives). Instead, participants wanted to experience an "ideal everyday," which involved two recurring subthemes: relaxation and third places.

Relaxation is self-explanatory: travelers, usually on leisure trips, preferred a slow-paced day with few responsibilities to a quick, busy day. A1 appreciated a relaxing or "chill" environment, as did B1, who elaborated that, as a busy professor, she often doesn't get a chance to do the "everyday" things that are part of this ideal day. She

gave an example of buying a birthday card for a friend: she plans to do this on an upcoming trip to visit friends, just because that will be the only time she has to do it.

Third places, such as bars, cafes, and bookstores as described in [33], are also a key part of this "ideal everyday." Many participants described local venues they loved: a coffeeshop and a taqueria (A13), cafes where he's seen friends he knows sitting outside (A14), cafes and dive bars (B4). B1 went as far as to suggest that she would travel to a place based on where the best coffee shops were. Because third places tend to be neutral, accessible, status-leveling places, travelers appreciate them. Stepping into another place's everyday life involves adjustments, and these third places give travelers a way to recharge.

Dimension 5: Authenticity

One final recurring theme involved travelers' desires for an "authentic" "non-touristy" place. Clearly, "touristy" places have some disadvantages: they are expensive (B6 gave the example of paying £39 to see the Crown Jewels in London) and often people act differently there (B7 described feeling like she "had a dollar sign on her forehead" in the tourist beaches of Cancún). But those inconveniences don't explain the intensity of the desire to be "not a tourist" (or even "the anti-tourist", as A9 described himself). Furthermore, some people appreciated touristy places, for practical reasons: B7 noted that not speaking Spanish limited her experience in Mexico, and A6 described how she would search for a place that's not the #1 tourist destination but also not completely local, due to language issues.

To understand this touristiness tension, it is useful to review previous work about authenticity in tourist places. Early work located all spaces on a 6-stage scale from front-stage (purely for show) to backstage (fully authentic) and predicted that all tourists would seek authenticity [24]. Later work added more nuance, describing the "authenticity" of an experience in nine subtypes depending on how authentic the place was, how authentic the people were, and whether the visitor put importance on the authenticity of the people or the place, both, or neither [34]. Furthermore, the authenticity of an experience may be best explained as existential authenticity, or the personal resonance with that experience. Existential authenticity has two forms: intra-personal (discovering and being true to oneself) and inter-personal (having a real connection to others) [48].

Different people may enjoy a trip to the Van Gogh Museum in Amsterdam for many reasons. They may appreciate seeing the original *Sunflowers* (objective authenticity) or seeing the official, definitive collection of Van Gogh's art (constructive authenticity). They may enjoy a stirring resonance with Van Gogh's masterful brushstrokes, or the ability to discuss these paintings with their fellow tourists (existential authenticity, intra- and inter-personal respectively). They might get useful information from the docents in an official, front-stage capacity, or they might get a docent to reveal little-known backstage stories about working at the museum. Finally, afterwards, they may stay in the enclavic tourist "bubble" of the Museumplein outside, or they may head to a more heterogeneous neighborhood, as

described in [9]. Each of these experiences may be regarded by one person as "authentic" and by another as "touristy."

This aspect of tourism, more even than the other four, is therefore impossible to definitively characterize or measure, but by a consideration of the people involved and the different types of authenticity, we can hope to provide guidance to get people to experiences that will resonate authentically for them.

Finding 3: Current search tools do not adequately investigate those five dimensions

Given that these five neighborhood characteristics (safety and room for everyone, aesthetics, serendipity, the ideal everyday, and authenticity) matter in different ways for different searchers, how do travelers search for neighborhoods now?

The primary search method used was to ask friends and family. If people visited friends, like B2 in Albuquerque and Portland, they can do this directly; otherwise, like A9, they would ask friends beforehand what were interesting and fun neighborhoods.

Online research was also widely used, often as simply as searching Google for "things to do in (city)" or "London off the beaten path" (B6). B7 lamented, though, that this kind of searching can turn the usually-fun process of traveling into work.

Because searching was so labor intensive, some people who did not have any pre-existing heuristics (as described in Finding 1) tried to create their own heuristics. A11 would search for the "queerest neighborhood" in a given city, as she did when she visited Zurich. This was not in order to find particular sites there (Zurich's queerest neighborhood featured one main gay bar and one main sex shop, neither of which she visited), but just because she found that she would often like the kind of people she met there. Similarly, B1 searched for the best coffee shops, not because she would spend most of her time there, but because she usually likes neighborhoods that have good coffee shops. B2 would read books about a place, like Gregory David Roberts's novel *Shantaram* before visiting Mumbai, or Maya Angelou before visiting San Francisco, in order to recognize places they mentioned.

When given the printed material about these places, participants agreed that they could be useful, but there were many caveats. Statistics would be helpful, but would need context, especially for unfamiliar numbers like density. Yelp and other point-oriented tools are helpful, but do not directly solve users' problems. Tweets, as in the selected words from the Twitter Neighborhoods TF-IDF Map, were usually disregarded. Finally, photos were tricky: some thought that they perfectly reflected their neighborhood, like A11. But some thought the opposite: A13 said that if he had seen the photos of his neighborhood, he might not have moved there, though he likes it now. A11 also mentioned that sometimes photos represent a neighborhood coincidentally: an octopus sculpture in her neighborhood was one example, but if it had been picked in a nearby neighborhood, it would have poorly reflected it.

As a result, we see an opportunity for a higher level of abstraction. The abstraction that participants liked the most was comparison to neighborhoods in cities that they know. This is similar to work that has been done both in research [22] and in popular culture [39]. Because people already know what neighborhoods in their own city are like, this can give them an easy way to understand neighborhoods in a new city. I will elaborate on how we will do this more in the next chapter.

Chapter 4: Proposed Work

From related research, we know that people are traveling in new ways and want to experience different things when they travel. From some related work and some of my prior work, I know that public geotagged social media posts can be an accurate and useful window into the culture of a neighborhood. From introductory interviews, I have identified details about the dimensions people want to explore when they travel: Safety and Room, Aesthetics, Serendipity, Ideal Everyday, and Authenticity. They don't need to "maximize" these dimensions, because these dimensions are complicated and individual, but should be able to browse them.

To help them accomplish this, I plan to build a web-based neighborhood guide. This will involve two parts: neighborhood comparison and context. Users will first be prompted to provide a city they know, a city they are traveling to, and a neighborhood to use as a basis for comparison.

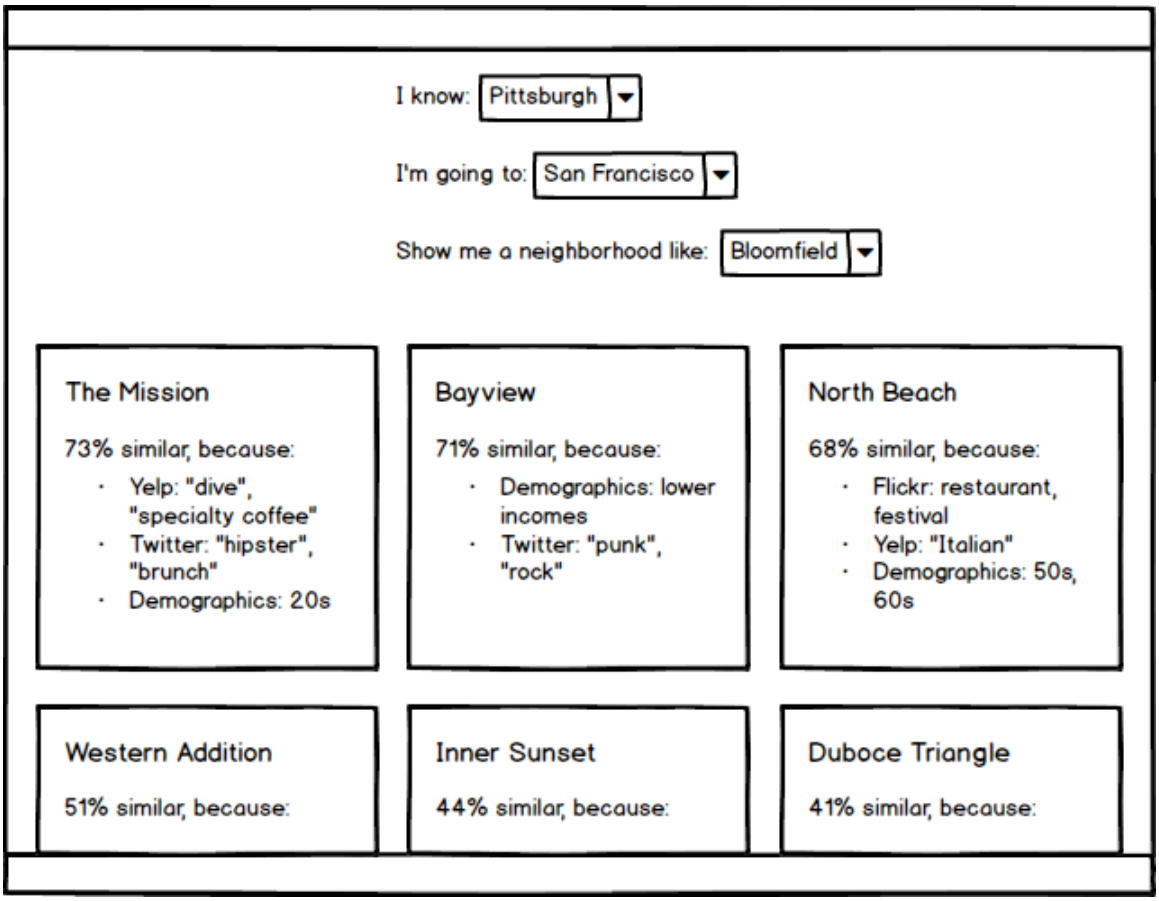

Figure 4: Mockup of the proposed neighborhood guide

I will use neighborhood comparison as the central metaphor because, in our interviews, I found that it was the most compelling metaphor to guide people's

neighborhood searches. People usually already know about neighborhoods in their own city, so I can use that knowledge to scaffold their process of learning about a new city.

However, providing a comparison ("The Williamsburg of Pittsburgh is Lawrenceville") is not enough. Research suggests that unintelligible systems can cause lack of trust and acceptance [23], and my interviewees echoed that concern: "I'd like [neighborhood comparisons], but I don't know if I could rely on it" (B3). Therefore, I must develop an algorithm that can easily be broken down, so the site can explain *why* Lawrenceville is the Williamsburg of Pittsburgh.

In the rest of this chapter, I will explain the neighborhood comparison algorithm I will implement, the ways I will add context to explain the algorithm's findings, and the evaluations I plan to run.

Neighborhood Comparison Algorithm

In introductory interviews, I found five main dimensions that people used in order to understand neighborhoods, so I will base the neighborhood comparison algorithm on those five dimensions. Each dimension will yield a feature vector. We can compare two feature vectors using a measure such as cosine similarity to find a similarity score between 0 and 1, and adding the five scores will give us a similarity score for any pair of neighborhoods. In this section, I will describe how we will find the feature vectors for each dimension.

Safety and Room for Everyone: US Census Demographics and Crime Statistics

This dimension is relatively straightforward because travelers in our original study preferred both safety and diversity. From the US Census, I plan to extract the percent of local residents who fit into each decade age group, income brackets, and racial breakdowns. I will also find crime statistics, in terms of crimes per person per year and violent crimes per person per year.

Aesthetics: Flickr autotags

Gathering data on aesthetic characteristics of neighborhoods is a more complicated endeavor, but for this we can turn to Flickr. Flickr photos have computer vision-based "autotags" attached to them, which identify the objects seen in the image (such as "people" or "sunset"). As a result, we can use the publicly available YFCC100M dataset [45], which contains about 49 million geotagged photos, to find photos in each neighborhood, then identify how many times each autotag shows up in a given neighborhood. This will give us a 1720-element feature vector, as Flickr currently recognizes 1720 distinct autotags.

Preliminary analysis suggests that these tags will show some difference in character between different neighborhoods. For example, in San Francisco, the 10 most common autotags in the Financial District are:

```
architecture, people, building, face, blackandwhite,
vehicle, monochrome, building complex, road, text
```

while the 10 most common autotags in the Outer Sunset (a residential/beach neighborhood) are:

```
nature, people, landscape, face, shore, seaside, sky, road,
water, coast
```

Serendipity: Walkscore and Transitscore

As explained in the previous chapter, the opportunity for serendipity in a place depends in a large part on how easy it is to get around by walking. Therefore, the serendipity feature vector will consist of two components: the previously developed walkability and transit-friendliness scores from Walkscore[5]. (Walkscore also provides a bikeability score, but as most travelers do not have bicycles, we expect this would be less helpful.)

Ideal Everyday: Third Places from Yelp

Because many travelers seem to want to experience the "Ideal everyday", including a relaxed pace and plenty of third places like cafés and bars, I plan to use Yelp reviews of these third places to capture how people describe a place. Users' star ratings are not particularly descriptive (nor are they trustworthy, as A1, A3, A9, and A10 independently mentioned), but the reviews contain rich descriptions. These descriptions will give a user an idea whether to expect upscale cocktail bars or greasy spoon diners, which will give them some sense of what the Ideal Everyday in this neighborhood is like.

However, these reviews are unstructured text. I see two promising options to turn this unstructured text into feature vectors:

1. Build frequency counts of all the words in all of these reviews throughout the city, then compare to a standard word frequency distribution in a news or Wikipedia corpus, in order to identify the most frequent words in Yelp reviews, compared to the language as a whole. Then use a bag-of-words approach for each neighborhood's venues to get frequency counts of each word in each neighborhood.
2. Use doc2vec as implemented in the Gensim library [40], which is based on the Paragraph Vector algorithm [21]. Paragraph Vector may be able to outperform bag-of-words models on tasks like this because it maintains some of the structure of the related sentence.

I plan to implement both of these and use whichever yields better results.

A third option, if these options prove problematic, is simply to create a vector of types of third places. For example, Bloomfield has 2 cocktail bars, 4 dive bars, 2 pubs, 1 gay bar, 0 sports bars, 5 coffee shops, and 1 tea room. This would likely not

[5] https://www.walkscore.com/

be as rich an information source to work with (because classifications are always incomplete and lacking nuance) but it would be a simpler option in case options 1 and 2 fail.

Authenticity: Tweets

As described in Chapter 3: Completed Work, authenticity is something that travelers want, but that is hard to define. Given that anyone may have a different definition of "touristy" or "authentic", perhaps the most value we can create here is by reflecting what has been said publicly on Twitter. As such, the tweets in the neighborhood become documents, and we can turn them into feature vectors in the same way as the Third Places reviews above.

Because authenticity is so individual, giving people a sense of what people in the neighborhood are saying is the best way we can give them a sense of whether they would resonate with that neighborhood. There's no way to learn and predict the "most authentic" neighborhood, because that designation is subjective enough to be meaningless. Someone wanting an "authentic" old-fashioned Pittsburgh experience might consider Shadyside an inauthentic yuppie neighborhood and visit the Strip District instead, while someone else might consider the Strip District an inauthentic tourist attraction and visit Shadyside to see what the "real" people in Pittsburgh do.

Again, for each dimension, once we have a feature vector, we can compute its similarity to other neighborhoods' feature vectors for that dimension. Given that we will only have on the order of 50-100 neighborhoods for any given pair of cities, we can compute these similarity values exhaustively; we do not need to use any more efficient nearest-neighbor algorithm. The computation of similarity between two example neighborhoods is illustrated in Figure 5. Note that, while I will start with a simple arithmetic mean of the five similarity values, I will adjust this algorithm based on user feedback, as I will explain in the next section.

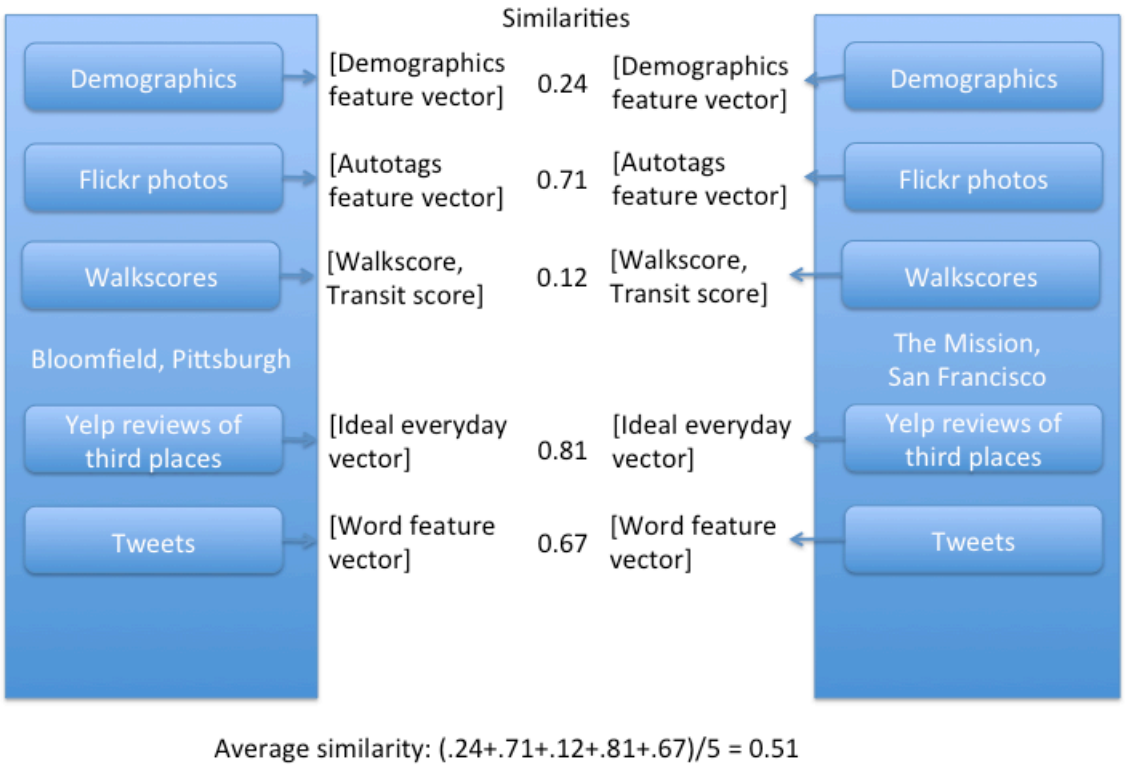

Figure 5: Illustration of similarity computation between two neighborhoods

This algorithm extends prior work in neighborhood comparison [22], but I want to emphasize the difference in the approach. While [22] used only Foursquare checkins to compare the venues in different neighborhoods, I am using far more types of social media data to develop a much richer comparison. Also, this prior work used neighborhood characterizations like "the student neighborhood" or "the fancy shopping district" while asking people to create a labeled data set; I will not use any such *a priori* labels. This will enable us to characterize neighborhoods that do not have a simple description, and include all of the nuance that comes from sources beyond lists of the venues in a place.

Context for Neighborhood Comparisons

It is important to be able to explain why neighborhoods are similar, so context will be an integral part of this application. As in Figure 4, I plan to show, with each similarity prediction, an indicator of why those two neighborhoods are similar. Given that we have five distinct dimensions, which all predict a similarity value between 0 and 1, it's easy to tell which dimension contributed most to the similarity rating. In the example in Figure 5, the vectors from the Yelp reviews of third places were the most similar, so it would be easy to report "Bloomfield and the Mission have a similarity rating of 0.51, mostly because the bars and cafés are the most similar."

However, we can give even more context than that. For each dimension, we can provide further context. For the demographics, we can show graphs of why the demographics of the neighborhoods are similar. If the Flickr photo autotags are the most similar, we could show which tags caused this similarity, and show example photos with those tags. We can use a photo summary such as those in [1, 15] to best show representative photos. If the Walk and Transit scores are the most similar, we can show the Walkscore and TransitScore maps through an embedded Walkscore map. If the Yelp reviews or Tweets are the most similar, we can surface which key phrases or words cause that similarity.

Evaluation

This web site aims to make people's trips better, so the gold standard study to evaluate its usefulness would be to have people make a trip without using the site, then make a trip with it, and evaluate their enjoyment of each trip. A study this large would be outside the scope of this thesis, but there are subsets of the application that can be evaluated and improved.

Are the neighborhood comparisons "right"?

Prior work [22] has approached neighborhood comparison as a classification problem, built a data set of "ground truth" neighborhood comparisons from a user study, and measured prediction accuracy. However, unlike many prediction tasks, it is hard to say what the "right" answer for a neighborhood comparison is. Is Lawrenceville really the Williamsburg of Pittsburgh? If someone argues that East Liberty is instead, there is really no way to prove either viewpoint right or wrong. Therefore, I will focus not on being "right," but on being close.

To evaluate this, I will run an online user study, in which I recruit people who know two different cities, and serve them neighborhood comparison predictions in one of three ways:

- At random (baseline)
- Using demographics and counts of venues only
- Using the five-part algorithm described in this chapter, which includes demographics but also social media posts

We will then ask them to evaluate if each comparison is plausible (see Figure 6). Assuming that our predictions exceed the baselines, this will provide evidence that social media aids in our understanding of cities and neighborhoods, at least when viewed through the traveler's lens.

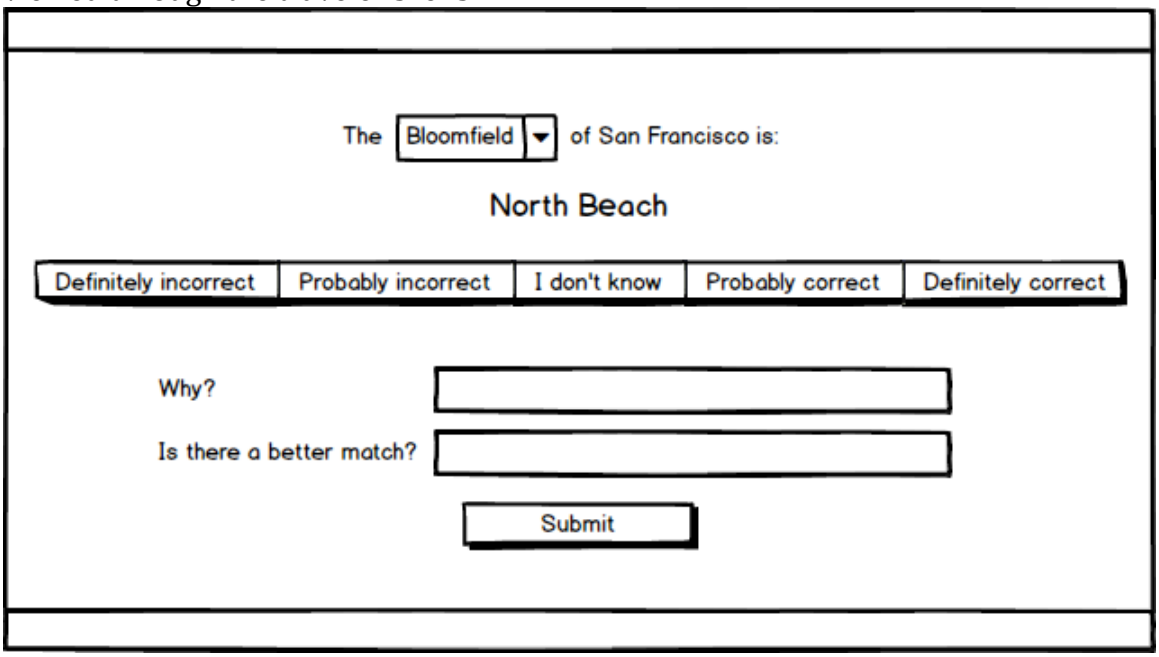

Figure 6: User interface mockup for the neighborhood comparison accuracy evaluation task

I will recruit participants online, from Craigslist, Reddit, and other forms of social media. I will recruit these people one city at a time, focusing on Pittsburgh and San Francisco, so that I can easily describe the request. ("Help us compare Pittsburgh's neighborhoods to other cities'" is easier to understand than "Help us compare any two cities.") I hope to recruit 50 people per condition, so 150 total. I will restrict recruitment to people who have lived in Pittsburgh or San Francisco and another city for at least 6 months.

Is this guide useful? Does it reflect the city accurately?

This will be more difficult to evaluate, but as it is more important, I want to at least try. I will run two user studies with people who are about to go on a trip, simply trying the application out. I will investigate both what parts of it they find most useful and how else they gather information.
This will help further develop the five dimensional model: verify that the dimensions I've chosen are important, understand if there are more dimensions, and learn more about why they find those dimensions important.

For one of these studies, I will recruit participants from among MHCI students in March, traveling to Austin for the SXSW conference, because there will be a lot of them, so we can get diverse data about one particular comparison of cities. I will also recruit people as they become available throughout the year.

Contributions

In developing these guides, I will make a working website that helps travelers find neighborhoods they will enjoy. This work will lead to the following research contributions:

- A model of tourist information search, focusing on five primary characteristics that tourists deem valuable today, based on formative interviews and qualitative insights from user studies.
- The iterative design and implementation of an automatically generated web-based neighborhood guide, which uses social media to provide comparisons between neighborhoods in different cities and to provide context for these comparisons.
- A deeper understanding of how social media can represent neighborhoods, based on the development and iterative feedback on this guide.

Chapter 5: Schedule

Early to mid May 2016: conference presentations at CHI and ICWSM, write CSCW paper based on introductory interviews.
Late May-June 2016: investigate algorithms for word embedding in vector spaces to determine which method to use. Develop scrapers to download Yelp reviews and Flickr autotags.
July-August 2016: wedding and honeymoon to China
Late August-September 2016: create a preliminary/skeleton web site, in order to be able to use it as a test bed.
October 2016: include data from two cities in the site. Recommend neighborhoods using a preliminary/skeleton algorithm.
November 2016: run an initial qualitative user study
December 2016: continue development on site
January 2017: run the quantitative user study comparing three different comparison algorithms (random, demographics only, and demographics plus social media)
February 2017: further continue site development
March 2017: run second qualitative user study with MHCI students going to SXSW.
April-May 2017: write thesis, defend in May

Acknowledgements

Jennifer Chou helped with the construction of the Twitter Neighborhoods TF-IDF Map web app and developing our TF-IDF algorithm. Alex Sciuto helped with analyzing data and paper writing for the paper “Our House, In The Middle Of Our Tweets.”

References

1. Ahern, S., Naaman, M., Nair, R., and Yang, J.H.-I. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. Proceedings of the 2007 conference on Digital libraries - JCDL ’07, (2007), 1.
2. Ashworth, G., & Page, S. J. (2011). Urban tourism research: Recent progress and current paradoxes. Tourism Management, 32(1), 1–15. http://doi.org/10.1016/j.tourman.2010.02.002
3. Bock, K. (2015). The changing nature of city tourism and its possible implications for the future of cities. European Journal of Futures Research, 3(1), 20. http://doi.org/10.1007/s40309-015-0078-5
4. Buck, M., Ruetz, D., & Freitag, R. (2014). ITB World Travel Trends Report.

5. Cheng, Z., Caverlee, J., Lee, K., & Sui, D. (2011). Exploring Millions of Footprints in Location Sharing Services. ICWSM. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2783/3292
6. Crandall, D., Backstrom, L., Huttenlocher, D., and Kleinberg, J. Mapping the World's Photos. Proceedings of the 18th International Conference on World Wide Web, Madrid, (2009), 761–770.
7. Cranshaw, J., Schwartz, R., Hong, J.I., and Sadeh, N. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. ICWSM, (2012).
8. Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). Social media update 2014. Pew Research Center, (January), 18. http://doi.org/10.1111/j.1083-6101.2007.00393.x
9. Edensor, T. (2001). Performing tourism, staging tourism. Tourist Studies, 1(1), 59–81. http://doi.org/10.1177/146879760100100104
10. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In KDD (Vol. 2, pp. 635–654). http://doi.org/10.1.1.71.1980
11. Füller, H., & Michel, B. (2014). "Stop Being a Tourist!" New Dynamics of Urban Tourism in Berlin-Kreuzberg. International Journal of Urban and Regional Research, 38(4), 1304–1318. http://doi.org/10.1111/1468-2427.12124
12. Gao, Y., Tang, J., Hong, R., Dai, Q., Chua, T.-S., & Jain, R. (2010). W2Go: a travel guidance system by automatic landmark ranking. Proceedings of the International Conference on Multimedia - MM '10, 123. http://doi.org/10.1145/1873951.1873970
13. Hao, Q., Cai, R., Wang, C., et al. Equip Tourists with Knowledge Mined from Travelogues. Proc. of the 19th International World Wide Web Conference, (2010), 1–10.
14. Horozov, T., Narasimhan, N., & Vasudevan, V. (2006). Using location for personalized POI recommendations in mobile environments. Proceedings - 2006 International Symposium on Applications and the Internet, SAINT 2006, 2006, 124–129. http://doi.org/10.1109/SAINT.2006.55
15. Jaffe, A., Naaman, M., Tassa, T., and Davis, M. Generating summaries and visualization for large collections of geo-referenced photographs. Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06, (2006), 89.
16. Kafsi, M., Cramer, H., Thomee, B., and Shamma, D. a. Describing and Understanding Neighborhood Characteristics through Online Social Media. WWW, (2015).
17. Kennedy, L., Naaman, M., Ahern, S., Nair, R., and Rattenbury, T. How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections Categories and Subject Descriptors. ACM Multimedia, (2007).
18. Krumm, J. (2007). Inference Attacks on Location Tracks. Pervasive Computing, 10(Pervasive), 127–143. http://doi.org/10.1007/978-3-540-72037-9_8

19. Krumm, J. and Horvitz, E. Eyewitness : Identifying Local Events via Space-Time Signals in Twitter Feeds. Proceedings of the 23nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, (2015).
20. Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2010). Travel route recommendation using geotags in photo sharing sites. Proc. 19th ACM International Conference on Information and Knowledge Management, 579–588. http://doi.org/10.1145/1871437.1871513
21. Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. International Conference on Machine Learning - ICML 2014, 32, 1188–1196. Retrieved from http://arxiv.org/abs/1405.4053
22. Le Falher, G., Gionis, A., & Mathioudakis, M. (2015). Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In ICWSM.
23. Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09, 2119–2129. http://doi.org/10.1145/1518701.1519023
24. MacCannell, D. (1977). Staged Authenticity: arrangements of Social Space in Tourist Settings. American Journal of Sociology, 682(3), 678–682.
25. Mahmud, J., Nichols, J., & Drews, C. (2013). Home Location Identification of Twitter Users. ACM Transactions on Intelligent Systems and Technology. Retrieved from http://tist.acm.org/papers/TIST-2012-11-0192.R1.pdf
26. Mahmud, J., Nichols, J., & Drews, C. (2012). Where Is This Tweet From? Inferring Home Locations of Twitter Users. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (pp. 511–514).
27. Maitland, R. (2010). Everyday life as a creative experience in cities. International Journal of Culture Tourism and Hospitality Research, 4(3), 176–185. http://doi.org/10.1108/17506181011067574
28. Maitland, R. (2013). Backstage Behaviour in the Global City: Tourists and the Search for the “Real London.” Procedia - Social and Behavioral Sciences, 105(0), 12–19. http://doi.org/10.1016/j.sbspro.2013.11.002
29. McNaught, C. and Lam, P. Using wordle as a supplementary research tool. Qualitative Report 15, 3 (2010), 630–643.
30. Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter's firehose. Proceedings of ICWSM, 400–408. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPDFInterstitial/6071/6379
31. Mummidi, L. N., & Krumm, J. (2008). Discovering points of interest from users’ map annotations. GeoJournal, 72(3-4), 215–227. http://doi.org/10.1007/s10708-008-9181-5
32. Okuyama, K., & Yanai, K. (2013). A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the Web. The Era of Interactive Media, 657–670. http://doi.org/10.1007/978-1-4614-3501-3_54

33. Oldenburg, R. (1989) The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community. Da Capo Press, Boston.
34. Pearce, P. L., & Moscardo, G. M. (1986). The Concept of Authenticity in Tourist Experiences. Journal of Sociology, 22(1), 121–132. http://doi.org/10.1177/144078338602200107
35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=2078195
36. Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., & Almeida, V. (2012). We Know Where You Live: Privacy Characterization of Foursquare Behavior. Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12, 898. http://doi.org/10.1145/2370216.2370419
37. Priedhorsky, R., Culotta, A., & Valle, S. Y. Del. (2014). Inferring the Origin Locations of Tweets with Quantitative Confidence. In CSCW (pp. 1523–1536).
38. Rattenbury, T., Good, N., and Naaman, M. Towards automatic extraction of event and place semantics from flickr tags. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07, (2007), 103.
39. Read, M. (2014). This Is the Williamsburg of Your City: A Map of Hip America. Retrieved January 1, 2016, from gawker.com/this-is-the-williamsburg-of-your-city-a-map-of-hip-ame-1460243062
40. Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50.
41. Richards, G. (2010). Tourism Development Trajectories - From Culture to Creativity? Encontros Científicos - Tourism & Management Studies, (6), 9–15. http://doi.org/10.4324/9780203933695
42. Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In Proceedings 1996 IEEE Symposium on Visual Languages (pp. 336–343). http://doi.org/10.1109/VL.1996.545307
43. Stors, N., & Kagermeier, A. (2015). Motives for using Airbnb in metropolitan tourism – why do people sleep in the bed of a stranger? Regions Magazine, 299(1), 17–19. http://doi.org/10.1080/13673882.2015.11500081
44. Takeuchi, Y., & Sugimoto, M. (2006). CityVoyager : An Outdoor Recommendation System Based on User Location History. Ubiquitous Intelligence and Computing, 4159(Figure 1), 625–636. http://doi.org/10.1007/11833529_64
45. Thomee, B., Shamma, D. a., Friedland, G., Elizalde, B., Ni, K., Poland, D., … Li, L. (2015). The New Data and New Challenges in Multimedia. arXiv Preprint arXiv:1503.01817, 1–7. http://doi.org/10.1145/2812802
46. Toyama, K., Logan, R., & Roseway, A. (2003). Geographic location tags on digital images. Proceedings of the Eleventh ACM International Conference on Multimedia - MULTIMEDIA '03, (November), 156. http://doi.org/10.1145/957044.957046

47. Wakamiya, S., Lee, R., and Sumiya, K. Crowd-sourced Cartography: Measuring Socio-cognitive Distance for Urban Areas based on Crowd's Movement. Proceedings of the 2012 ACM Conference on Ubiquitous Computing, (2012), 935–942.
48. Wang, N. (1999). Rethinking authenticity in tourism experience. Annals of Tourism Research, 26(2), 349–370. http://doi.org/10.1016/S0160-7383(98)00103-0
49. Yannopoulou, N., Moufahim, M., & Bian, X. (2013). User-Generated Brands and Social Media: Couchsurfing and Airbnb. Contemporary Management Research, 9(1), 85–90. http://doi.org/10.7903/cmr.11116
50. Yelp Wordmap. http://www.yelp.com/wordmap/sf, 2016. http://www.yelp.com/wordmap/sf.
51. Zhang, A.X., Noulas, A., Scellato, S., and Mascolo, C. Hoodsquare: Modeling and Recommending Neighborhoods in Location-based Social Networks. SocialCom, (2013), 1–15.